OPTIMAL SPARSE KERNEL LEARNING FOR HYPERSPECTRAL ANOMALY DETECTION

Prudhvi Gurram, Heesung Kwon

U.S. Army Research Laboratory
ATTN: RDRL-SES-E
2800 Powder Mill RD
Adelphi, Maryland 20783

Zhimin Peng, Wotao Yin

Rice University
Department of Computational
and Applied Mathematics
Houston, Texas 77005

ABSTRACT

In this paper, a novel framework of sparse kernel learning for Support Vector Data Description (SVDD) based anomaly detection is presented. In this work, optimal sparse feature selection for anomaly detection is first modeled as a Mixed Integer Programming (MIP) problem. Due to the prohibitively high computational complexity of the MIP, it is relaxed into a Quadratically Constrained Linear Programming (QCLP) problem. The QCLP problem can then be practically solved by using an iterative optimization method, in which multiple subsets of features are iteratively found as opposed to a single subset. The QCLP-based iterative optimization problem is solved in a finite space called the *Empirical Kernel Feature Space* (EKFS) instead of in the input space or *Reproducing Kernel Hilbert Space* (RKHS). This is possible because of the fact that the geometrical properties of the EKFS and the corresponding RKHS remain the same. Now, an explicit nonlinear exploitation of the data in a finite EKFS is achievable, which results in optimal feature ranking. Experimental results based on a hyperspectral image show that the proposed method can provide improved performance over the current state-of-the-art techniques.

Index Terms— Sparse kernel learning, Optimal feature selection, Empirical kernel feature space, Empirical kernel map

1. INTRODUCTION

Feature selection for learning algorithms aims to find a relevant subset of features that can improve the learning performance by discarding features not useful or even harmful for the given tasks. In the case of kernel-based anomaly detection, such as SVDD, the feature selection requires the accurate estimation of the contribution of each feature to the objective function, i.e., the radius of a hypersphere in the RKHS.

In this paper, a new framework of optimal sparse kernel learning for SVDD-based anomaly detection (OSKLAD) is proposed. The proposed OSKLAD optimally extends the feature selection technique used for the kernel-based learning approaches [1] into SVDD-based anomaly detection by fully optimizing the feature selection method for nonlinear kernels in a newly defined finite space called the EKFS [2]. Hence, the OSKLAD can be considered as a fully optimized version of the wrapper approach to the SVDD-based anomaly detection with nonlinear kernels. The initial objective of the proposed OSKLAD begins with finding a single subset of original features that can be used to build an optimal hypersphere in the RKHS. This objective can be modeled as a Mixed Integer Programming (MIP) problem. However, the MIP problem is NP-hard, and so the MIP model is relaxed into a Quadratically Constrained Linear Programming (QCLP) problem [3] by converting the objective function of the MIP problem into lower bounded quadratic inequality constraints. This QCLP problem is yet intractable due to the prohibitively large number of the inequality constraints. To address this issue, a cutting plane method based on the *restricted master problem* coupled with Multiple Kernel Leaning (MKL) [4] is iteratively used. The goal is to find only a small subset of the inequality constraints that are actively used to define the feasible region of the parameters of the given QCLP problem.

The *active* constraints are effectively identified by finding the *most violating constraints* instead whose half-planes maximally violate the corresponding inequality constraints. Therefore, the task becomes finding multiple subsets of *most violated features* associated with the corresponding *most violating constraints* given the objective function, such as the radius of a hypersphere in the RKHS. However, finding the most violating constraints also becomes a combinatorial problem, if nonlinear kernels, such as Gaussian RBF kernel or high order polynomial kernels, are used, due to the prohibitively large number of possible combinations (subsets) of the original features. To tackle this issue, in the proposed OSKLAD, the most violated features are found in the EKFS. The EKFS is a finite space linearly spanned by basis vectors, which are generated by a map, called the Empirical Kernel Map (EKM). It is shown that the EKHS and the corresponding RKHS constructed by using the same kernel function have the same geometrical property. This means that solutions of any optimization problem obtained from either space are identical. In the proposed OSKLAD, the subsets of the most violated fea-

tures are optimally found in the EKFS since individual feature ranking in terms of contribution to the radius in the EKFS can be performed optimally based on the property of canonical dot product and the finite dimensionality of the space.

2. OPTIMAL SPARSE KERNEL LEARNING

In this section, we present an optimal sparse kernel learning for anomaly detection (OSKLAD) using SVDD as a basic building block. Inspired by the feature selection approach for the kernel-based classification [1], the OSKLAD addresses the problem of the optimal feature selection for the SVDD-based anomaly detection. The basic formulation of OSKLAD is to minimize the radius of the enclosing hypersphere while allowing outliers except that in OSKLAD, only a subset of features is used. So, the model is described as a mixed integer programming problem:

$$\begin{aligned} \min_{\mathbf{d}} \min_{R,\xi_i,\mathbf{a}} \; & R^2 + C \cdot \sum_{i=1}^{N} \xi_i \\ \text{subject to } & \|\Phi(\tilde{\mathbf{x}}_i) - \mathbf{a}\|^2 \leq R^2 + \xi_i \\ & \xi_i \geq 0 \\ & \tilde{\mathbf{x}}_i = \mathbf{x}_i \odot \mathbf{d}, \; i = 1, 2, ..., N, \end{aligned} \tag{1}$$

where $\mathbf{d} \in \mathbb{D} = \{\mathbf{d} | d_j \in \{0,1\}, \sum_{j=1}^{M} \mathbf{d}_j = B, j = 1, 2, ..., M\}$, and $\odot$ represents elementwise product. Here B is a threshold that controls the number of features that are selected. If one assumes that $\mathbf{d}$ is fixed in Eq. 1, it turns into a continuous constrained optimization problem just like a standard SVDD. By applying the Langrange multipliers and KKT conditions to it, we can derive the dual problem (similar to standard SVDD) as:

$$\begin{aligned} \min_{\mathbf{d}} \max_{\alpha_i} & \sum_{i=1}^{N} \alpha_i k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_i) - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \\ \text{subject to} & \sum_{i=1}^{N} \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C \\ & \tilde{\mathbf{x}}_i = \mathbf{x}_i \odot \mathbf{d}, \; i = 1, 2, ..., N. \end{aligned} \tag{2}$$

However, one should notice that Eq. 2 is still a mixed interger programming (MIP) problem due to the last constraint, which is computationally expensive to solve. In order to solve this problem, it can be converted into a Quadratically Constrained Linear Programming (QCLP). We define $S(\alpha, \mathbf{d}) = \sum_{i=1}^{N} \alpha_i k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_i) - \sum_{i,j=1}^{N} \alpha_i \alpha_j k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$, and introduce an additional parameter t to obtain the QCLP equivalent of (2) as follows:

$$\begin{aligned} \max_{\mathbf{a},t} \; & t \\ \text{subject to } & \sum_{i=1}^{N} \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C \\ & t \leq S(\alpha, \mathbf{d}), \; \forall \, \mathbf{d} \in D. \end{aligned} \tag{3}$$

Though Eq. 3 is convex, a large number of inequality constraints (last condition in Eq. 3) makes it impractical to be solved by existing techniques. The number becomes huge if the features reside in a high dimensional space. Note that not all the inequality constraints used in Eq. 3 are actively used in defining the feasible region of the optimization problem. In fact, only a small number of the constraints are useful and directly used to solve the optimization problem. Therefore, an iterative algorithm can be used, in which instead of solving Eq. 3 at once, an intermediate solution pair (t, α) is iteratively updated based on a limited subset of previously found active constraints. This optimization problem is called the *restricted master problem*, which is closely related to the cutting plane algorithm described in [5]. The *restricted master problem* consists of two steps [6]: 1) (t, α) are optimized based on a previously found restricted subset $\mathcal{I}$ of features, which maximally violates the constraints; and 2) a new vector $\mathbf{d}$ of the most violated features is obtained based on newly optimized (t, α) in step 1 and added to the restricted subset $\mathcal{I} = \mathcal{I} \bigcup \mathbf{d}$. These two steps are iterated until convergence [7]. Finding $\mathbf{d}$ of the most violated features is detailed in the next subsection.

The intermediate solution pair (t, α) is now obtained from the following optimization problem

$$\begin{aligned} \max_{\mathbf{a},t} \; & t \\ \text{subject to } & \sum_{i=1}^{N} \alpha_i = 1, \\ & 0 \leq \alpha_i \leq C, \\ & t \leq S(\alpha, \mathbf{d}^l), \; \mathbf{d}^l \in \mathcal{I}. \end{aligned} \tag{4}$$

Let $\mu_l \geq 0$ be the dual variable for each constraint in Eq. 4. The Lagrangian of Eq. 4 can be written as:

$$L(t, \mu) = t + \sum_{l=1}^{p} \mu_l S(\alpha, \mathbf{d}^l). \tag{5}$$

By setting $\frac{\partial L}{\partial t} = 0$, we have $\sum_{l=1}^{p} \mu_l = 1$. The Lagrangian $L(t, \mu)$, after applying this partial KKT condition, can be rewritten as $L(t, \mu) = \sum_{l=1}^{p} \mu_l S(\alpha, \mathbf{d}^l)$, which transforms

(3) to the following problem:

$$\begin{aligned} \max_{\alpha} \min_{\mu} \ & \sum_{l=1}^{p} \mu_l S(\alpha, \mathbf{d}^l) \\ \text{subject to } & \sum_{i=1}^{N} \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C \text{ for } i = 1, 2..., N \\ & \sum_{l=1}^{p} \mu_l = 1, \mu_l \geq 0 \text{ for } l = 1, 2..., p. \end{aligned} \tag{6}$$

One can observe that can be solved using a two-step iterative process to obtain optimal sparse weights of individual kernels μ and optimal lagrange multipliers α^* (which define the support vectors or the enclosing hypersphere).

3. OPTIMAL FEATURE SELECTION: FINDING MAXIMALLY VIOLATING FEATURES

For updating $\mathbf{d}$, the features that maximally violate the last constraint in Eq. 3 need to be determined. Since the goal of Eq. 3 is to maximize t, and it is upper-bounded by $S(\alpha, \mathbf{d})$ according to the constraint, the features that maximally violate this constraint will minimize $S(\alpha, \mathbf{d})$. One has to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{d}} \ & S(\alpha, \mathbf{d}) \\ \text{subject to } & \sum_{i=1}^{M} d_i = B \\ & d_i \in \{0, 1\}. \end{aligned} \tag{7}$$

In this section, we describe the method to find these feature vectors for both linear kernel and non-linear kernel.

3.1. Linear Kernel

If a linear kernel is used, since $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$, we have $S(\alpha, \mathbf{d}) = \sum_{j=1}^{M} d_j c_j$, where $c_j = \sum_{i=1}^{N} \alpha_i x_{ij}^2 + (\sum_{i=1}^{N} \alpha_i x_{ij})^2$. $S(\alpha, \mathbf{d})$ is a linear function of $\mathbf{d}$. Once we have optimal support vectors, the global solution of $\mathbf{d}$ can be easily obtained by sorting c_j's in ascending order and setting the first B corresponding elements in $\mathbf{d}$, d_j to 1 and the rest to 0. Once the optimal feature subset is chosen for a kernel, optimal α and μ are updated by solving Eq. 6. These two steps are repeated until the algorithm converges.

3.2. Non-linear Kernel

If a Gaussian RBF kernel is used, $S(\alpha, \mathbf{d})$ is not a linear function of $\mathbf{d}$. We cannot solve the problem in Eq. 7 optimally because of the large number of combinations of features that have to be considered. So, the data is tranformed from infinte dimensional RKHS into another space called *empirical kernel feature space* (EKFS) with finite dimensionality using *empirical kernel map* (EKM). This will allow us to select subsets of features optimally while still preserving the nonlinear correlations among the features. For a given set of training data points $\{\mathbf{x}_i\}_{i=1}^n$, the map defined by

$$\Phi_n : \mathbb{R}^n \to \mathbb{R}^n \text{ where } \mathbf{x} \mapsto k(\cdot, \mathbf{x}) = (k(\mathbf{x}_1, \mathbf{x}), \ldots, k(\mathbf{x}_n, \mathbf{x}))^T \tag{8}$$

is called the EKM with respect to $\{\mathbf{x}_i\}_{i=1}^n$ [2]. However, the kernel function k used to build kernel matrices in previous subsections cannot be represented using Φ_n, since they do not form an orthonormal system. The dot product to use in the representation of k is the not the canonical dot product in the EKFS $\mathbb{R}^n$. In order to turn Φ_n into a feature map associated with k, EKFS is endowed with a dot product $\langle \cdot, \cdot \rangle_n$ such that $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi_n(\mathbf{x}_i), \Phi_n(\mathbf{x}_j) \rangle_n$. After analyzing certain conditions using this equality as shown in [2], the dot product $\langle \cdot, \cdot \rangle_n$ can be converted to a canonical dot product by merely whitening the EKFS and using the new basis functions as features. It can be represented as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi_n^w(\mathbf{x}_i), \Phi_n^w(\mathbf{x}_j) \rangle, \tag{9}$$

where the feature map in whitened EKFS is given by

$$\Phi_n^w : \mathbf{x} \mapsto K^{-\frac{1}{2}} (k(\mathbf{x}_1, \mathbf{x}), \ldots, k(\mathbf{x}_n, \mathbf{x}))^T. \tag{10}$$

where K is the Gram matrix and $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. The kernel function in Eq. 9 is used to build the kernel matrices in Eqs. 2-7. Hence, the feature subset selection problem turns exactly into (7) (linear version) except for the fact that in this case the features are selected in EKFS. Similar to the OSKLAD with a linear kernel, the overall Optimal Sparse Kernel Learning for Anomaly Detection (OSKLAD) in the EKFS is described in Algorithm 1.

Algorithm 1 OSKLAD with nonlinear kernel

1: Map the data points into the EKFS by using a certain kernel k
2: Initialized: $\alpha = \frac{1}{N}\mathbf{1}$, find the maximally violating feature subset $\mathbf{d}$, and set $\mathcal{I} = \{\mathbf{d}\}$.
3: Run SKAD based on the kernel matrices generated by $\mathcal{I}$ and optimize for α and μ.
4: Find the next maximally violated feature subset $\mathbf{d}$ based on the current α and μ and set $\mathcal{I} = \mathcal{I} \bigcup \{d\}$.
5: Repeat steps 3-4 until convergence.

4. SIMULATION RESULTS

In this section, the performance of OSKLAD is evaluated on a hyperspectral digital imagery collection experiment(HYDICE) image, which contains 30 small painted pannels located in the background. We chose a small patch

(69 pixels $\times$ 10 pixels) as the background data set, which is used to obtain the radius R and the center of the hypersphere. The distance of each test pixel in the image to the center of the hypersphere is determined. If the distance is greater than R, the pixel is considered as an anomaly, otherwise, it is a background pixel. In our experiments, the performance of SVDD, SKAD [8][1] and OSKLAD with both linear and Gaussian RBF kernels are compared with one another. For SVDD and SKAD, both linear and Gaussian RBF kernel are used in the input space. For OSKLAD with linear kernel, feature selection is performed in the input space. However, for OSKLAD with Gaussian RBF kernel, the input vector is first mapped into EKFS using EKM. At this point, we can just use linear kernel in EKFS, which translates to using Gaussian RBF kernel in the input space as described in the previous sections. The kernel bandwidth parameter is determined by implementing the minimax technique on randomly selected 10 regions of the image to represent the background as done in [9]. The same value is used over all the test pixels in the image for all the algorithms.

The number of features used for each hypersphere of SKAD with both linear and Gaussian RBF kernel and OSKLAD with linear kernel is 75, which is half of the total number of features. For OSKLAD in the EKFS, the total number of features available after mapping the pixels from input space to EKFS is reduced to 96, and we have used 48 features for each hypersphere. Fig.1 shows the anomaly detection results for SVDD, SKAD and OSKLAD with both linear and Gaussian RBF kernels. The value of each pixel in the results is the ratio of the distance between the pixel and the radius of the hypersphere. For comparison, we normalize the scaled in all the resulting images to be between 0 and 1. One can see that all the six methods are able to identify the first two rows of anomalies, but OSKLAD in EKFS can identifies anomalies with much less noise (clean background) and it is also able to detect the small targets in the third row.

5. CONCLUSIONS

In the proposed work, to achieve optimality in kernel-based feature selection for anomaly detection using SVDD, the QCLP problem is optimally solved in a new finite space called the Empirical Kernel Feature Space (EKFS) instead of the RKHS. Experimental result show that by optimally selecting features, significant improvements can be made in hyperspectral anomaly detection in EKFS rather than the original input space.

6. REFERENCES

[1] M. Tan, L. Wang, and I. W. Tsang, "Learning sparse SVM for feature selection on very high dimensional datasets," in *ICML*, Haifa, Israel, June 2010, pp. 1047–1054.

[2] B. Schölkopf and A. J. Smola, *Learning with Kernels*, The MIT Press, Massachusetts, 2002.

[3] R. Hettich and K. O. Kortanek, "Semi-infinite programming: Theory, methods, and applications," *SIAM Review*, vol. 35, no. 3, pp. 380–429, Sept. 1993.

[4] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "Simplemkl," *J. Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

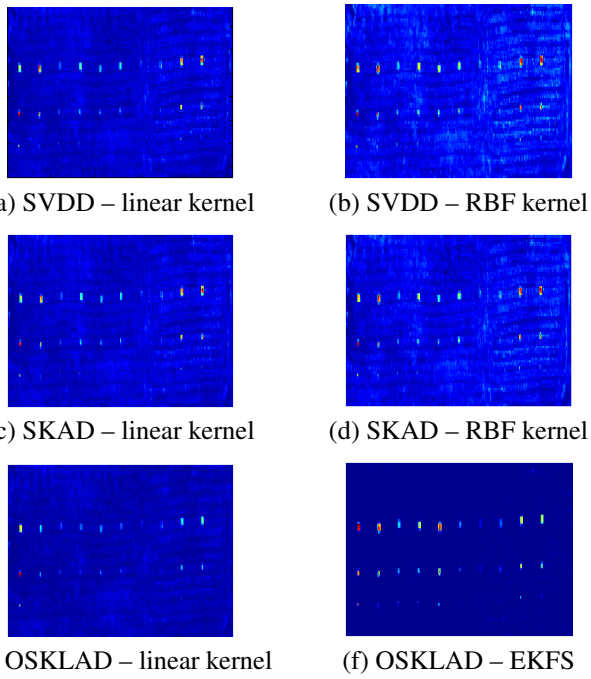
[5] J. E. Kelly, "The cutting plane method for solving convex problems," *J. Soc. Indust. Appl. Math.*, vol. 8, no. 4, Dec. 1960.

[6] J. Chen and J. Ye, "Training SVM with indefinite kernels," in *ICML*, Helsinki, June 2008, pp. 136–143.

[7] P. Gurram, Z. Peng, H. Kwon, and W. Yin, "Optimal sparse kernel learning for anomaly detection," *Pattern Recognition*, under review.

[8] P. Gurram, H. Kwon, and T. Han, "Sparse kernel-based hyperspectral anomaly detection," vol. 9, no. 5, pp. 943–947, Sept. 2012.

[9] Amit Banerjee, Philippe Burlina, and Chris Diehl, "A support vector method for anomaly detection in hyperspectral imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 8, pp. 2282–2291, 2006.



(a) SVDD – linear kernel (b) SVDD – RBF kernel

(c) SKAD – linear kernel (d) SKAD – RBF kernel

(e) OSKLAD – linear kernel (f) OSKLAD – EKFS

Fig. 1. Anomaly detection results of the HYDICE image using SVDD, SKAD and OSKLAD

[1] Sparse kernel-based anomaly detection (SKAD) has been developed by two of the current authors